The ISL Statistical Translation System for Spoken Language Translation

Stephan Vogel[+]*, Sanjika Hewavitharana*[+]*, Muntsin Kolss*[*]*, Alex Waibel*[+*]*,*

Interactive Systems Laboratories
[+]Carnegie Mellon University, Pittsburgh, USA
[*]University of Karlsruhe, Karlsruhe, Germany
[vogel+, sanjika, ahw]@cs.cmu.edu, muntsin@ira.uka.de

Abstract

In this paper we describe the components of our statistical machine translation system used for the spoken language translation evaluation campaign. This system is based on phrase-to-phrase translations extracted from a bilingual corpus. A new phrase alignment approaches will be introduced, which finds the target phrase by optimizing the overall word-to-word alignment for the sentence pair under the constraint that words within the source phrase are only aligned to words within the target phrase. The system will be used for Chinese-to-English translations under the small, additional and unlimited data conditions, and for the small Japanese-to-English translation track.

1. Introduction

Statistical machine translation (SMT) is currently the most promising approach, esp. to large vocabulary text translation. In the spirit of the Candide system developed in the early 90s at IBM [Brown et al. 1993], a number of statistical machine translation systems have been presented in the last few years [Wang and Waibel 1998], [Och and Ney 2000], [Yamada and Knight 2000]. These systems share the basic underlying principles of applying a translation model to capture the lexical and word reordering relationships between two languages, complemented by a target language model to drive the search process through translation model hypotheses. Their primary differences lie in the structure and source of their translation models. Whereas the original IBM system was based on purely word-based translation models, current SMT systems try to incorporate more complex structure.

The statistical machine translation system developed in the Interactive Systems Laboratories (ISL) uses phrase-to-phrase translations as the primary building blocks to capture local context information, leading to better lexical choice and more reliable local reordering. A new approach to extract phrase translation pairs from bilingual data has been developed, which is not using the Viterbi alignment, but is based on optimizing a constraint word-to-word alignment for the entire sentence pair. This is described in Section 2.1.

Finding good phrase translation pairs is very important. But as a source phrase can have alternative translations, it is also necessary to assign meaningful probabilities to those alternatives. Typically, longer phrases are seen only a few times. Probabilities estimated from relative frequencies are therefore not reliable. We therefore calculate phrase translation probabilities based on the word-to-word translation probabilities, as described in Section 2.3.

Section 3 outlines the architecture of the decoder that combines the translation and language model to generate complete translations.

The BTEC corpus is a very limited domain corpus and therefore many test sentences are close to one or several sentences seen in the training data. We implemented and tested a simple translation memory component, which will be described in Section 4.

Finally, in Section 5 we present a series of experiments in the Chinese-to-English and Japanese-to-English translation tasks. The Basic Travel Expression Corpus (BTEC) is used as domain-specific data [Takezawa et al. 2002]. Different data conditions are explored: small in-domain data only, using additional out-of-domain data, and using a larger in-domain corpus.

2. The Models

2.1. Phrase Alignment

The ISL translation system uses word-to-word and phrase-to-phrase translations, extracted from the bilingual corpus. Different phrase alignment methods have been explored in the past, like extracting phrase translation pairs from the Viterbi path of a word alignment, or simultaneously splitting source and target sentence into phrases and aligning them in an integrated way [Zhang 2003]. For the experiments reported in this paper a new phrase alignment method was explored.

2.2. Phrase Alignment via Constrained Sentence Alignment

Assume we are searching for a good translation for one source phrase $\tilde{f} = f_1...f_k$, and that we find a sentence in the bilingual corpus, which contains this phrase. We are now interested in finding a sequence of words $\tilde{e} = e_1...e_l$ in the target sentence, which is an optimal translation of the source phrase. Any sequence of words in the target sentence is a translation candidate, but most of them will not be considered translations of the source phrase at all, whereas some can be considered as partially correct translations, and a small number of candidates will be considered acceptable or good translations. We want to find these good candidates.

The IBM1 word alignment model aligns each source word to all target words with varying probabilities. Typically, only one or two words will have a high alignment probability, which for the IBM1 model is just the lexicon probability. We now modify the IBM1 alignment model by not summing the lexicon probabilities of all target words, but by restricting this summation in the following way:

- for words inside the source phrase we sum only over the probabilities for words inside the target phrase candidate, and for words outside of the source phrase we sum only over the probabilities for the words outside the target phrase candidates;
- the position alignment probability, which for the standard IBM1 alignment is $1/I$, where I is the number of words in the target sentence, is modified to $1/(l)$ inside the source phrase and to $1/(I-l)$ outside the source phrase.

More formally, we calculate the constrained alignment probability:

$$p_{i_1,i_2}(f|e) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} p(f_j|e_i) \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} p(f_j|e_i) \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} p(f_j|e_i)$$

and optimize over the target side boundaries i_1 and i_2.

$$(i_1, i_2) = \underset{i1,i2}{\text{argmax}}\{p_{i_1,i_2}(f|e)\}$$

It is well know that 'looking from both sides' is better than calculating the alignment only in one direction, as the word alignment models are asymmetric with respect to aligning one to many words. Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words:

$$p_{i_1,i_2}(e|f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1...j_2)} p(e_i|f_j) \times \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} p(e_i|f_j) \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1...j_2)} p(e_i|f_j)$$

To find the optimal target phrase we interpolate both alignment probabilities and take the pair (i_1, i_2) which gives the highest probability.

$$(i_1, i_2) = \underset{i_1,i_2}{\text{argmax}}\{(1-c)p_{(i_1,i_2)}(f|e) + cp_{(i_1,i_2)}(f|e)\}$$

Actually, we take not only the best translation candidate, but all candidates, which are within a given margin to the best one. All candidates are then used in the decoder, when also the language model is available to score the translations. The phrase pairs can be either extracted from the bilingual corpus at decoding time or stored and reused during system tuning. It should also be mentioned that single source words are treated in the same way, i.e. just as phrases of length 1. The target translation can then be one or several words.

2.3. Phrase Translation Probabilities

Most phrase pairs $(\tilde{f}, \tilde{e}) = (f_{j_1}...f_{j_2}, e_{i_1}...e_{i_2})$ are seen only a few times, even in very large corpora. Therefore, probabilities based on occurrence counts have little discriminative power. In our system we calculate phrase translation probabilities based on a statistical lexicon, i.e. on the word translation probabilities ($p(f, e)$):

$$p(\tilde{f}|\tilde{e}) = \prod_j \sum_i p(f_j|e_i).$$

2.4. The Language Model

The language model used in the decoder is a standard 3-gram language model. We use the SRI language model toolkit [SRI-LM Toolkit] to build language models of different sizes, using the target side of the bilingual data only or using additional monolingual data.

2.5. Position Alignment Model

Different languages have different word order. In the standard word alignment models this is captured by word position models, e.g. absolute positions $p(i|j, I, J)$ in IBM2 alignment model or relative positions $p(i|i_{prev}, I)$ in the HMM alignment model [Vogel et al. 1996]. We use a simplified relative position model in our SMT decoder.

$$p(i|i_{prev}, I) = e^{-\frac{|i-i_{prev}|}{c}} \quad (1)$$

with a suitably chosen constant c. This constant is essentially a scaling factor for the model when combining it with the other models in the decoder.

2.6. Sentence Length Model

Source sentence and target sentence are typically of different length. However, when using a large bilingual corpus to collect the sentence length statistics, it becomes clear that the probability distribution $p(I|J)$, where J is the number of words in the source sentence and I is the number of words in the target sentence, is rather flat and therefore does not seem to be very helpful. On the other side we observe that the language model typically prefers shorter translation. To compensate for this we use a simple sentence length model, which gives a constant bonus for each word generated. Putting a higher weight on the sentence length model contribution to the overall translation score results in generating translations, which are on average longer.

3. Decoding

Statistical machine translation is based on the noisy channel approach:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f) = \underset{e}{\operatorname{argmax}}\, p(f|e)p(e) \qquad (2)$$

The components are the language model $p(e)$, for which we use a trigram language model, and the translation model $p(f|e)$, which in our case is composed of the word and phrase translations. The argmax denotes the search algorithm, which finds the best target sentence given those models. Applying the language model requires that the previous words are known. This leads to a search organization which constructs the target sentence in a sequential way. However, to incorporate the different word order of different languages the words in the source sentence have to be covered non-sequentially while the translation is generated.

In the current implementation we allow for phrase-to-phrase translation. Decoding proceeds essentially along the source sentence. At each step, however, the next word or phrase to be translated may be selected starting from all positions within a given look-ahead window from the current position. The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations and, if available, other specific information like named entity translation tables are used to build a translation lattice. This lattice contains the all partial translations as building blocks, from which the complete translation has to be generated.

A standard n-gram language model is then applied to find the best path in this lattice. It is during this search that reordering has to be taken into account, by jumping ahead a few positions, filling in the gap later on. To ensure full coverage of the source sentence each partial translation carries information about the source words already translated.

Standard pruning strategies are employed to keep decoding time within reasonable bounds. The ISL decoder allows for flexible pruning, as the language model history, the translated position and the number of generate words can be used individually and in combination in pruning. Details have been described in [Vogel et al. 2003] and especially in [Vogel 2003].

4. A Translation Memory Component

The BTEC data consists of typical phrases used in the tourism and medical domain. The sentences are usually short, on average only 6-7 words, and many have similar patterns, as shown here with Spanish-English sentence pairs from the BTEC corpus:

en qué tipo de trabajo estás interesado ? what kind of job are you interested in ?
en qué tipo de cosas estás interesado ? what kind of things are you interested in ?
en qué tipo de excursiones estás interesado ? what kind of tour are you interested in ?

Fiven a test sentence we will often find the same or a very similar sentence in the training corpus. For the 506 sentences in the Chinese-English development test set) 5% or the test sentences were identical to a sentence in the training corpus, 20% of the sentences could be matched with one insertion, deletion, or substitution error only, and another 24% matched with 2 errors. For the close matching sentences the idea is to start from the given translation and to make some simple corrections.

The translation memory works as follows: For each test sentence $S_f = f_1...f_J$ we find the closest matching source sentence $S'_f = f'_1...f'_{J'}$ in the training corpus. The similarity is measured in terms of edit distance. The translation of S'_f, which is $S'_e = e_1...e'_I$ is also extracted.

If there is an exact match, we output Se' as the desired translation of S_f. For those sentences with error 1, we decide what type of operation (substitution, deletion or insertion) is required to produce the correct translation. Also identify the words f' in S'_f and f in S_f that has to be altered. The repair operations allow for multi-word substitutions, deletions, and insertions on the target side.

Depending on the type of the operation needed, one of the following operations is performed.

1. Substitution of f' with f:

   i. Find all possible phrase alignments e' in S'_e for the word f'.

   ii. Find all possible translations e of word f.

   iii. Replace e' with e to produce S_e.

   iv. Score the resulting translation (S_f, S_e) with the translation and language model.

   v. Iterate over all e' and e and choose the best S_e as the desired translation.

2. Deletion of f' in S'_f:

   i. Find the possible phrase alignments e' in S'_e for the word f'.

   ii. Remove e' from S'_e to produce S_e.

   iii. Score the resulting translation (S_f, S_e) with the translation and language model.

   iv. Iterate over all e' and choose the best S_e as the desired translation.

3. Insertion of word f into S'_f:

   i. Find all possible translations e for word f.

   ii. Insert e into a position i in S'_e to produce S_e.

   iii. Score the resulting translation (S_f, S_e) with the translation and language model.

   iv. Iterate over all translations e and all word positions i in S'_e and choose the best S_e as the desired translation.

To find the target phrase which needs to be repaired, or candidate translations used in the repair operations, the phrase alignment method described in SectionPhrases was used.

To integrate the results from SMT and the Translation Memory we simply replaced the SMT translation of close matching sentences, with the translation produced by translation memory approach.

5. Experiments

Experiments where performed to study the effect of different training data conditions. As in-domain data the BTEC corpus was used, a corpus created at ATR [Takezawa et al. 2002] and extended with translations into different languages by the CSTAR partners. In the small data track, only a part of the BTEC corpus was used. The so-called additional data track allowed for bilingual and monolingual data available from LDC. In the unrestricted data track the full BTEC corpus could be used.

5.1. Evaluation

We report translation results using the well known Bleu [Papineni 2001] and NIST mteval [MTeval 2002] scores. The the NIST mteval script version 11a was used to calculate both the NIST and the Bleu score. One peculiar feature of the Bleu metric implementation in the NIST mteval v011a script is the calculation of the reference length, which is used to calculate the length penalty. Whereas the original implementation sums the length of the reference translation, which is closest to the length of the system translation, the NIST implementation sums over the length of the shortest reference translation. This leads to very different length penalties in the two metrics. For the Chinese data the reference length for NIST is 3601.7 words, whereas the reference length for Bleu is 2429 words, i.e. about one third shorter. This has, of course, a big effect on the tuning of the system: translations scoring high on the Bleu metric will be much shorter than translations getting high NIST scores.[1]

5.2. The Test Data

Results are reported for Chinese-to-English and Japanese-to-English translation tasks. Two test sets were used for each language: one development test set (Dev), which was used to tune the parameters of the translation system and a test set (Test), which was translated using the optimal parameter settings. All test sets were provided by ATR with word segmentation. For evaluation 16 reference translations were used, whereby not all references were created as genuine translations, but as paraphrases. Table 1 gives the details for all four test sets.

Table 1: Translation results for the Chinese small data track.

	Chinese		Japanese	
	Dev	Test	Dev	Test
Sentences	506	500	506	500
Words	3515	3794	4108	4370
Vocabulary	870	893	954	979

The number of unknown words differ depending on the training data and will be given in each case below.

5.3. Chinese Small Data Track

The Chinese small data track uses 20,000 sentence pairs, where the Chinese sentences are already word segmented. It has to be assumed that the word segmentation of the training data matches the word segmentation of the test data. In the next sub-section we will see that word segmentation makes a difference and that higher translation quality can be achieved with re-segmenting both training and test data.

Table 2 gives the details for the data used in the Chinese small data track evaluation.

Different setups for the translation system were tested. Results are given in Table 3. First, the IBM1 lexicons $p(f_j|e_i)$ and $p(e_i|f_j)$ were used in the phrase alignment step, but the translation probability for the phrase pairs was estimated from the relative frequencies. Next, the phrase translation probability was calculated using the

[1] This difference in implementation for the calculation of the length penalty has been pointed out to Mark Przybocki, the implementor of the current mteval version, and also a number of researcher using this script, but it was not considered to be a significant problem. It is clear that this problem arises only with several reference translations and is esp. severe when the test sentences and therefore the reference translations are very short, as is the case with the BTEC data.

Table 2: Training and test data statistics Chinese small data track.

	CH	EN
Sentences	20,000	
Words	182,902	188,935
Vocabulary	7,645	7,181
LM PP	–	68.6
Unk in Dev	160	–
Unk in Test	104	–

IBM1 lexicon and the HMM lexicon respectively. Each time we see an improvement in translation quality, both when tuned towards high Bleu scores and when tuned towards high NIST scores. Finally, n-best list rescoring with the HMM lexicon was gave a small improvement in Bleu score, but none in NIST score. An improvement of about 1.8 in Bleu score and 0.24 in NIST score is statistically significant on the 95% level. That is to that that the improvements from using relative frequencies to using the IBM1 lexicon for scoring the phrase translations, and then again using the HMM lexicon leads to a statistically significant improvement in Bleu score. For NIST score the step from using the IBM1 lexicon to using the HMM lexicon is statistically significant.

Table 3: Translation results for the Chinese small data track.

	Opt. Bleu		Opt. NIST	
	Bleu	NIST	Bleu	NIST
IBM1 Lex, Rel Freq	41.6	4.69	36.5	7.58
IBM1 Lex, IBM1 Lex	43.5	6.07	39.5	7.67
HMM Lex, HMM Lex	46.0	5.77	36.8	7.94
- n-best rescoring	46.7	4.87	–	–

We tested the translation memory component for sentences which matched exactly or had only one error. There are 130 sentences in development set for which this condition holds. The parameter setting for the SMT system was set to generate translations, which where somewhat balanced with respect to NIST and Bleu score, leaning somewhat more towards a high NIST score. Replacing the 130 sentences, which were translated by the translation memory module, did not improve Bleu and NIST scores, as can be seen in Table 4.

Table 4: Effect of using the translation memory component for the Chinese small data track.

	Bleu	NIST
SMT alone	39.1	7.90
With TM	38.8	7.84

There is a small, but not significant drop in both scores. But when the translations of the two methods are compared, in many instances, the translation memory (TM) has produced better 'quality' translation.

Ref:	how much does it cost to send this to japan
SMT	please send this to japan how much is it
TM	what is the cost for sending this to japan
Ref	do i have to transfer to get there
SMT	i 'd like to change trains to get there
TM	do i have to change buses to get there
Ref	could you repeat that please
SMT	would you please say it again please
TM	would you say it again please
Ref	what is today 's date
SMT	what is today's number
TM	what 's the date today

For the unseen test data translation with parameter settings for High Bleu, High NIST, and a more balanced version were generated and evaluated. Results are given in Table 5. It turned out the the more balanced parameter setting gave a slightly higher NIST score than the parameter setting which gave highest NIST score on the development test set, and at the same time a much higher Bleu score. It can be assumed that the length ratio between source sentences and reference translations is somewhat different between the development and the test set.

Table 5: Translation results for the Chinese small data track on unseen test data.

	Bleu	NIST
High Bleu	44.6	7.31
High NIST	37.9	8.31
Balanced	41.4	8.34
With TM	36.7	8.16

5.4. Chinese Additional Data Track

In this data track additional data could be used to improve translation quality. However, this additional data was restricted to corpora which are distributed through LDC. All Chinese-English bilingual data available was therefore news data, which is to say, out-of-domain data. The question therefore is, if this data will improve translation quality, or rather harm it.

To use the additional data, first of all a resegmentation of the BTEC training corpus and also test data was necessary. Word segmentation is typically based on a word list and perhaps additional word frequency information. It is clear that using the vocabulary of the small BTEC corpus would not be helpful, as this word list is rather small and would not help to find an adequate segmentation of the news corpora. We therefore

applied the same word segmentation to the BTEC training and test data, which was also used to preprocess the additional LDC data. The word list used contains about 45,000 words. The statistics for the resulting corpus is shown in Table 6. It is interesting to notice that after re-segmenting the BTEC data the number of unknown words reduced significantly, from 160 to 89 for the development set and from 104 to 88 for the test set.

Table 6: Training and test data statistics Chinese additional data track.

	BTEC		3*BTEC+NEWS	
	CH	EN	CH	EN
Sentences	20,000		129,209	
Words	175,284	188,935	1,50m	1,65m
Vocabulary	7,617	7,181	25,961	32,658
LM PP	–	68.6	–	100.5
Unk in Dev	89	–	5	–
Unk in Test	88	–	13	–

To further reduce the number of unknown words, we can use the additional data. Adding just a large out-of-domain corpus will usually not help, but rather result in a degradation in translation quality. We therefore select from the large bilingual Chinese-English corpus only those sentences, which contain words and phrases occurring in the test data. More specific, for each n-gram in the test data, which occurs there k times, we select up to $10 * k$ sentences in the training corpus containing this n-gram. For the development and test set used in the experiments this resulted in a small corpus (NEWS) of about 1 million words, with a vocabulary of about 24K Chinese resp. 30K English words. This data was then added to the in-domain data and used to train translation and language model. To bias more towards the in-domain data we also trained models on a corpus, where the small BTEC corpus was added 3 times, the NEWS corpus only once.

The LM 3-gram perplexity for the 1+1 combination was 106.7, whereas for the 3+1 combination it was 100.5, compared to the 68.6, when using only the in-domain data for building the language model. This increase in perplexity shows that adding additional data goes both ways: reducing the number of unknown words, but also increasing the perplexity of the models.

Translation results are shown in Table 7. The re-segmentation alone gave already higher Bleu and NIST scores. However, when adding the out-of-domain data the scores went down, indicating worse translation quality. Only after biasing the models more towards the in-domain data a small, yet statistically significant improvement could be achieved over using the in-domain data alone.

Again, three parameter setting where used to translate the unseen test sentences, using the system trained on the combined data, 3 times the BTEC corpus plus NEWS corpus once. Results are given in Table 8. When we compare these results with the small data track scores, then we see that both High Bleu score High NIST score are higher when adding the out-of-domain data. Again, these improvements are statistically significant.

Table 7: Translation results for the development test set in the Chinese additional data track.

	Opt. Bleu		Opt. NIST	
	Bleu	NIST	Bleu	NIST
Re-segmented	48.7	5.42	38.2	8.16
BTEC + 1m NEWS	44.7	5.06	41.1	6.88
3*BTEC + 1m NEWS	51.0	5.09	39.9	8.33

Table 8: Translation results for the unseen test data in the additional data track.

	Bleu	NIST
High Bleu	48.5	5.85
High NIST	40.1	8.82
Balanced	43.0	8.22

5.5. Chinese Unrestricted Data Track

This data condition imposes no restrictions on which data to use for training the translation and language models. The most valuable data is, of course, in domain data. As the BTEC corpus contains more than 160k sentence pairs, we can compare the effect of additional in-domain data to using the additional out-of-domain data. The corpus statistics for the BTEC corpus used in this experiment is given in Table 9.

The interesting numbers here are that the full BTEC corpus leads to fewer unknown words, but when adding the sampled news data, the number of unknown words is the same as in the additional data track.

The LM perplexity for the reference translations is, on average, higher than when using only the 20,000 sentences to build the LM, increasing from 68.6 to 72.0, despite eight time as many data. This again indicates that these reference translations have are more varied then when generating genuine translations. For the combined corpus the perplexity is now lower, as the larger BTEC corpus gives a stronger bias towards in-domain data.

Here, we see first of all that more in-domain data boosts translation quality. The Bleu score increased by 5 points, i.e. a 10% relative improvement, and the NIST score increased by 0.9, also a 10% relative improvement. An the other side, additional out-of-domain data did not help to improve translation quality. The benefit of having fewer unknown words is lost by moving out-of-domain

Table 9: Training and test data statistics Chinese unrestricted data track.

	BTEC		3*BTEC+NEWS	
	CH	EN	CH	EN
Sentences	161,307		553,130	
Words	1,13m	1,21m	4,36m	4,70m
Vocabulary	12,619	13,358	27,978	36,075
LM PP	–	72.0	–	95.1
Unk in Dev	48	–	5	–
Unk in Test	52	–	13	–

Table 10: Translation results for the development test set in the Chinese unrestricted data track.

	Opt. Bleu		Opt. NIST	
	Bleu	NIST	Bleu	NIST
BTEC	53.8	6.35	47.2	9.09
3*BTEC+NEWS	53.3	6.63	45.9	9.10

with the translation and language model. Perhaps reducing the additional corpus to just those few sentences, which contain words not seen in the in-domain training data could help.

Table 11: Translation results for the unseen test data in the unrestricted data track .

	Bleu	NIST
High Bleu	57.1	7.60
High NIST	48.6	9.66
Balanced	52.5	9.56
With TM	51.3	9.29

5.6. Japanese Small Data Track - An Exercise in Language Portability

For the Japanese small data track the essential question was how fast good translation could be generated, given that a system for Chinese-to-English, which had similar characteristics in terms of corpus and vocabulary size, had already been build and tuned. So, the two IBM1 lexicons were trained and the language model from the 20k English sentences was built. The data could be used without additional preprocessing. Training the models is a matter of minutes. Therefore, the overall effort was rather small; formatting the reference translations for automatic evaluation was probably the most time consuming part.

The first translation runs used the parameter setting which gave highest Bleu and NIST scores for the Chinese small data track situation, when using the IBM1 lexicons for phrase pair extraction and phrase pair scoring. Additional tuning was then performed to see how close the initial translation was already to optimal performance. The results are given in Table 12. We see that the first translation gave already close to optimal results. Overall the effort to train and tune the Japanese-English translation system was less then half a day.

Table 12: Translation results for the Japanese small data track development test set, using parameters from optimal Chinese-English translation, and further optimizing for Japanese-English.

	Opt. Bleu		Opt. NIST	
	Bleu	NIST	Bleu	NIST
With CE Parameters	48.8	7.07	45.4	9.27
Additional Tuning	50.2	7.38	45.8	9.29

In Table **??** the results for the unseen test set are given. Results are somewhat lower than the scores obtained on the development data.

Table 13: Translation results for Japanese-English small data track on unseen test data.

	Bleu	NIST
High Bleu	46.3	6.73
High NIST	41.5	8.84
Balanced	43.0	8.06

6. Summary and Future Work

A new phrase alignment approach has be developed, which is based on finding for a given source phrase the target phrase by optimizing the alignment probability for the entire sentence pair under the restriction that words inside the source phrase align only to word inside the target phrase and words outside of the source phrase align only to words outside of the target phrase. Comparison with previously developed phrase alignment methods has shown that this new approach leads to comparable and even better results, and yet is very simple. A major advantage of this method is that with even using only an IBM1 lexicon, i.e. using only the simplest alignment model, which has the shortest training time, competitive results are possible. It seems likely that other co-occurrence statistics like Dice coefficient, Chi-square or mutual information might lead to similar results. On the other side, however, better lexicons do lead to better phrase alignment and thereby to better translation results.

A further advantage is that phrases up to any length can be found when applying the phrase search and alignment during decoding time.

Future extension will include using higher order word alignment models, like the HMM alignment model or the IBM4 alignment model in the phrase alignment step.

The translation memory component used in this study was rather simple. There are a number of possibilities how this work could be extended: 1. Allowing more than one mismatch between test sentence and sentence in the training corpus, esp. for longer sentences. 2. Instead of selecting only one of the most similar sentences, selecting the n-best matches and iterate over all of them. 3. Using additional information, like parts of speech, to have a more discriminative matching between sentences. 5. Integrating SMT and translation memory results using better criteria than just on the number of errors.

The experiments presented in this paper have shown that out-of-domain data can be used to improve translation quality when only a small domain specific corpus is available.

A major problem became apparent in the evaluation with using multiple reference translations, which are not original translations, but at least in part paraphrases of original translations. This make the reference translations less typical as shown by the increased language model perplexity when training the language model on the full BTEC corpus. Also the wide variability in length of the multiple reference translations and the different calculation for the length penalty in Bleu and NIST score calculation results in rather low correlation between these to metrics, and thereby also to low correlation with human evaluation. We observed as typical behavior that the higher the Bleu score the lower the NIST score and vice versa.

7. References

[Brown et al. 1993] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[MTeval 2002] NIST MT Evaluation Kit Version 11a. Available at: http://www.nist.gov/speech/tests/mt/.

[Och and Ney 2000] Franz Josef Och and Hermann Ney. Improved Statistical Alignment Models. *Proceedings of ACL-00*, pp. 440-447, Hongkong, China.

[Papineni 2001] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. *Technical Report RC22176 (W0109-022)*, IBM Research Division, T. J. Watson Research Center.

[SRI-LM Toolkit] SRILM - The SRI Language Modeling Toolkit. SRI Speech Technology and Research Laboratory. http://www.speech.sri.com/projects/srilm/

[Takezawa et al. 2002] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto. Toward a broad-coverage bilingual corpus for speech translations of travel conversations in the real world. *Proc. of Third Int. Conf. on Language Resources and Evaluation (LREC)*, pp. 147-152, Las Palmas, Spain, May 2002.

[Vogel et al. 1996] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation. in *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pp. 836–841, Copenhagen, August 1996.

[Vogel et al. 2003] Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel. The CMU Statistical Translation System. *Proceedings of MT Summit IX*, New Orleans, LA, U.S.A., September 2003.

[Vogel 2003] Stephan Vogel. SMT Decoder Disected: Word Reordering *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.

[Wang and Waibel 1998] Yeyi Wang and Alex Waibel. Fast Decoding for Statistical Machine Translation. *Proc. ICSLP 98*, Vol. 6, pp. 2775-2778, Sidney, Australia, 1998.

[Yamada and Knight 2000] Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. in *Proc. of the 39th Annual Meeting of ACL*, Nancy, France, 2000.

[Zhang 2003] Ying Zhang, Stephan Vogel and Alex Waibel. Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.